VIRTUAL DATA PRODUCTS IN AN INTELLIGENT ARCHIVE

07/18/2003
Version 1.0

Mark Clausen
Northrop Grumman, TASC
4801 Stonecroft Blvd. Chantilly, VA 20151-3822

Christopher S. Lynnes
Code 902
NASA/GSFC
Greenbelt, MD 20771

ABSTRACT

A key characteristic of intelligent archives of the future is adaptability to changing science and application demands and capabilities. One way to incorporate adaptability within a modest cost envelop is to provide "virtual" data products, that is, data that are produced from the raw data on-demand, rather than routinely producing and storing data. The virtual data product strategy has the benefit of decreasing storage requirements, making data migration to new technology easier and less expensive, providing users with more precise selection of spatio-temporal areas or parameters and more options in algorithm selection. Aside from the basic mechanics of implementing on-demand processing, a number of issues arise with product virtualization, though they all have potential solution paths. The computing power needed for such processing is hoped to be achieved through revolutionary computing developments. Difficulties in quality assurance of on-demand products could be addressed by a process-and-discard strategy with stored statistics, or by processing low-resolution versions of the data *a priori*. Likewise quality control of the raw input data would benefit from either of these strategies, or from a robust quality check of the input data. The data migration problem is transmuted to a software porting issue as new computer architectures, operating systems or compilers are introduced. Checking cross-platform portability during algorithm acceptance can mitigate this issue, coupled with the quality assurance strategies mentioned above. In the end, the desirability of replacing stored routinely processed products with virtual products is likely to be driven by the relative advancement of computing versus storage technologies. Should the former advance beyond the latter, virtual products will become both more feasible and more desirable.

TABLE OF CONTENTS

INTRODUCTION

An ever-increasing amount of data is being returned from larger numbers of spacecraft with more sensors. This increase is beginning to tax the ability of data archives to keep up with the data processing, storage and distribution requirements. The problem is expected to increase as hyperspectral instruments become more prevalent, as it will become less feasible to produce and store, or indeed even define, all potentially useful products on a fixed schedule. The introduction of adaptive intelligent data systems may add to the challenge, as they request new measurements or data in response to features or results just discovered in previous steps. An effort at NASA is underway to define the attributes and potential architectural concepts of intelligent archives in the future that could meet these challenges[1].

A critical component of such intelligent archives is clearly the ability to adapt to evolving scientific capabilities and demands; yet remain with a modest cost envelope. However, the strategy of routinely processing and storing products for later retrieval places severe constraints on a data management system: algorithm changes introduce inconsistencies in the time series; archive capacity planning must be years in advance to accommodate the space and power needs of large silos; data migrations to new technologies are costly and time-consuming, yet indispensable.

One possible approach to these challenges in an intelligent archive is the creation of "virtual" data products, *i.e.*, products which are not stored and then recovered from archive, but rather generated "on-demand" when requested. The concept of on-demand processing of remote sensing data has a long history. For example, Landsat products are typically processed on demand, as are those from the Earth Observing System's ASTER instrument. However, this type of on-demand processing is typically driven by requirements to vary run-time parameters at the time of processing. On-demand processing was also suggested as an alternative architecture for EOSDIS[2]. Ultimately, on-demand processing was not expanded beyond ASTER and Landsat, due to concerns about algorithm interdependence and product standardization[3]. On the other hand, only a few attempts have been made to substitute on-demand processing for storage of defined products, or in other words, storing "virtual" data products. One example is provided by the Grid Physics Network (GriPhyN) project, which has developed a prototype of a virtual data product system for the Laser Interferometer Gravitational-wave Observatory (LIGO)[4]. So far, their efforts have focused on constructing the infrastructure for product virtualization and materialization using standard Grid protocols. On the other hand, the University of Alabama—Huntsville has been working on the application of an online data mining system to provide virtual products[5]. These projects are developing the basic mechanics of virtual products, such as how to define and implement recipes for production or how to provide metadata for the virtual products. However, a number of issues beyond these basic mechanics remain if product storage is to be replaced by virtualization at any useful scale. Both the drivers for virtualization and the challenges are discussed in this white paper.

It is first important to understand how routine processing of remote sensing data for archives currently occurs. Firstly, it is useful to define different levels of data processing to describe the amount of processing accomplished to produce a data product. For the purposes of this paper, the definitions shown in Table 1 (based on the EOS data level definitions) are used.

Table 1 Data Level Definitions

Processing Level	Definition	EOSDIS Archive*	
		Million Granules	Volume in TB
Level 0	Instrument data at original resolution, time order-restored, with duplicate packets removed	1.6	412
Level 1A	Level 0 data that are reformatted with calibration data and other ancillary data included. Geolocation information for each spatial element (e.g., pixel) of the reformatted sensor-coordinate data is stored separately.	3.1	444
Level 1B	Level 1A data to which the radiometric calibration algorithms have been applied to produce radiances, irradiances, or brightness temperature.	9.3	663
Level 2	Geophysical parameter data retrieved from a single sensor's Level 1B data by application of geophysical parameter algorithms. The MODIS science team has an additional level, 2G, which contains pixel to grid mappings.	14.6	729
Level 3	Earth-gridded geophysical parameters that have been averaged, gridded, or otherwise rectified or composited in space and/or time.	5.5	175
Level 4	Model output or results of analyses from lower-level data; such as variables derived from data collected by multiple sensors	0.3	3

*Data archived from February, 2000 to July, 2003.

Although on-demand processing can be applied in principle to any level above level 0, the difficulty increases as the processing level increases. This is due to the increasing variety and volume of input data that must be staged, as well as the cumulative effect of the processing that must be done. Whereas Level 1 and Level 2 data, in swath or scene coordinates, typically need only the same scene for their upstream products, plus perhaps bounding scenes, Level 3 and Level 4 data typically require at least a full day's worth of data, and often much more depending on the algorithm. As a result, the cost/benefit equation for these higher levels makes them less attractive for virtualization; thus, we will focus on Level 1 and Level 2 data in our discussion. This is but a minor limitation, as the bulk of EOS archives, both in volume and number of granules, consists of Level 1 and Level 2 data (Table 1). The purpose of this paper is to examine the basic needs for product virtualization, develop a general concept of operations for its use, lay out a conceptual architecture and discuss some of the issues concerning its implementation.

THE BENEFITS OF PRODUCT VIRTUALIZATION

The section below discusses the drivers (or benefits) of product virtualization. The first two drivers, handling large amounts of data and mitigating migration difficulty, benefit primarily the cost containment of archive construction and management. However, the rest of the benefits are scientific in nature, relating to increasing the adaptability of the archive from the perspectives of both science algorithms and science users.

SUPPORT ARCHIVAL OF INCREASING AMOUNTS OF DATA

Increases in sensors, channels, spatial and temporal resolution, and derived parameters are all combining to drive archival requirements considerably upward. At some point, it becomes impractical, if not impossible (computationally and financially), to archive all derived parameters in addition to the raw Level 0 data. Furthermore, not all raw data need be processed to higher levels. Although most sensors now collect data globally, much of it may be of little scientific or practical interest. However, the higher-level data must be available on demand when needed. Virtualizing higher level data would allow only the data of interest, i.e., requested by some person or model, to be computed and distributed.

MITIGATE THE DIFFICULTY AND COSTS OF DATA MIGRATION

Data archives employing tape technology must usually migrate data every three to five years in order to keep the data in an archive that is still supported by the manufacturer. As data volume grows, this migration effort becomes increasingly expensive and time consuming[6]. One possible way of alleviating this would be to virtualize the Level 1 and Level 2 products and migrate only the irreproducible Level 0 and the relatively low-volume Level 3 and Level 4 products. Of course, this would transform the data migration problem from one of sheer volume to the problem of migrating data production algorithms to new platforms and operating system versions, a challenge which is addressed in a later section.

SUPPORT ADVANCES IN ALGORITHMS

The development of algorithms to derive geophysical parameters is a continuously evolving process. Rather than re-create derived data sets when new or updated algorithms are developed, virtual products should enable the algorithms, which are archived along with the Level 0 data, to be updated. Users of the virtual product archive can then access it to obtain the latest (and hopefully best) values for the derived Level 1 or Level 2 data when needed.

AVOID MULTIPLE VERSIONS OF DATA SETS

In Level 2 processing, a number of different algorithms for a given geophysical parameter are often current within the scientific community or are used to preserve continuity with previous instruments or algorithms. For example, Chlorophyll-a from the MODIS instrument is available using three (3) different algorithms[7]. Storing only the collected data and then deriving the desired parameters would ensure that users could have consistent (and latest) versions of the data. In cases where algorithm quality is debated, users could have the option to select which algorithm will drive the data reduction.

SIMPLIFICATION OF DATA COLLECTION FOR THE USER

With the large amounts of data and the potential for updated algorithms or multiple versions, data discovery and selection must be made simpler for users, especially for Level 2 products. Currently, when making requests for the various products available, users must often be knowledgeable about potential differences between various incarnations of

data sets. As the number of users from outside the remote sensing community increase, the process of choosing and obtaining these data must be made simpler.

A PRIORI SUBSETTING

In today's production concept, data "granules" of a fixed size, temporal coverage and spatial coverage are produced, often resulting in large data files with many parameters covering large geographic areas. Many users, however, want small segments of the data for just a few parameters. The result has been a significant effort to subset the data after the fact. An important benefit of virtualized production would be the ability to generate the desired parameters only over the user's study area.

CHALLENGES

Product virtualization faces several challenges in its development and operation. While many of these challenges might seem difficult or even insurmountable, in fact many of them can be addressed by the very intelligence in the Intelligent Archive that is driving the requirement for virtualization. The following sections discuss some of the more significant issues and potential areas of study to resolve them.

PROCESSING POWER

Perhaps the most obvious challenge to product virtualization is the potential need for processing power. We can expect the available computing power to continue to increase in silicon-based processors with Moore's Law for at least several years. Eventually, revolutionary computing technologies (*e.g.* quantum computing) may pick up the increase, albeit with a potential stagnant period in between. Another area that could enable the power for product virtualization is the Grid computing effort, which can use distributed computing power more flexibly, thus making more available for use. This could even be carried to the extreme of using the user's computer to produce the products by shipping the raw data, plus a portable version of the processing algorithm. The algorithm would then be automatically (or with one click) run on the user's computer, much the way self-extracting archives work today.

PROCESSING QUALITY CHECKS

Perhaps the greatest concern on the part of the end users is the ability to confirm the proper creation of the Level 1 or Level 2 data. This becomes a very real concern as computing improvements drive the migration of the intelligent archive to new hardware and/or software systems. One approach to this would be to compare quality checks created on the original system with those created on the new system in order to ensure a successful migration. Processing quality checks can also be employed by users to ensure the data have been properly processed; quality checks delivered with the returned data can be compared to those stored in the archive.

However, the kind of standard checksums (e.g. Cyclic Redundancy Checksums) currently applied by archives to ensure file integrity would not be adequate for ensuring information integrity of virtual products. This is because any two instantiations of the same virtual "granule" may contain production-related metadata (e.g. production time) which are unique to that instance, numeric roundoff differences, or byte alignment differences due to architecture. Thus, a more sophisticated means for assuring the quality of the underlying scientific values is required. Two potential methods could be applied to provide these checks.

STATISTICAL QUALITY CHECK

The first method would entail processing all Level 1 and Level 2 data upon receipt of the raw data. These data would be kept in a short-term archive for operational or quasi-operational use. Statistics (*e.g.*, mean, standard deviation, etc.) that describe the data over defined areas of sensor space (*e.g.*, over every few scan lines) would then be calculated. These descriptive statistics would then be stored as meta-data for long term archival in the intelligent archive along with the Level 0 data and the algorithms used to generate the higher-level products. When archived Level 0 data, quality checks, and algorithms are migrated to a new system (or when users request Level 1 or Level 2 data), the quality checks would be recalculated and compared to the original quality checks in the metadata.

One advantage of this method is that the quality values generated would (if wisely chosen) be a very small fraction of the space required for the overall archive. Additionally, Level 1 and Level 2 data would be immediately available for operational or quasi-operational users on a short-term archive. On the down side, this method requires processing of all data to Level 1 and 2 immediately, regardless of the need for these data in the short term. Furthermore, the acceptance of new processing algorithms would require a massive processing campaign to process all Level 1 and Level 2 data in order to generate new quality checks. However, it is likely that such processing would be required anyway in order to generate the Level 3 and Level 4 products. In this concept, the Level 3 and Level 4 products are produced and stored; the Level 1 and Level 2 products are discarded after saving statistical information for future quality checks in on-demand processing.

THUMBNAIL QUALITY CHECK

The second method would involve processing Level 1 and Level 2 data on some subset of the Level 0 data (*e.g.*, every 100^{th} pixel). (On the other hand, for some data products, alternative methods of generating low-resolution products might be needed.) Such "thumbnail" views would then be made available on a short-term archive as well as archived in the intelligent archive along with the Level 0 data and the algorithms used. When archived Level 0 data and algorithms are migrated to a new system (or when users request Level 1 or Level 2 data), the thumbnail values would be compared to the corresponding values in the full density data.

There are several advantages to this method. First, the thumbnail "images" could be hosted on a random access system for browsing; when the user selects data, the intelligent archive would "reach" into the long-term storage (i.e., a tape silo) for the archived Level 0 data and metadata. Second, the generation of these checks does not require processing of all Level 1 and Level 2 data. Finally, acceptance of new algorithms into the intelligent archive would likewise not require processing of all data. On the down side, the thumbnails might constitute a larger fraction of the overall archive than the statistical approach, depending on the resolution of the thumbnails.

QUALITY CONTROL OF "RAW" DATA

One of the lessons learned in current processing environments is that quality problems in the raw data are sometimes not detected until certain forms of higher level processing have been attempted. Thus, all virtual product algorithms would require a capability to detect quality problems in the input data. Alternatively, a robust quality control algorithm could be applied to the Level 0 data on arrival, or provided as an on-the-fly capability when retrieving for virtual products. However, there is also a risk that quality problems in the raw data would be discovered too late to do anything about it, such as reprocess frame data to level 0 or reacquire raw data from backup. This favors some detailed examination of the Level 0 data immediately, either via the robust QC method, or the process-and-discard method discussed in the previous subsection.

MIGRATION OF VIRTUAL DATA PRODUCTS

Any long-term archive must accommodate advancements in hardware and software technology over time, particularly in the area of operating system and compiler upgrades. Migrating large data sets to new hardware/software technologies has always been an issue with all large data sets. However, the product virtualization in an intelligent archive presents quite a different porting problem, that of migrating the algorithms to new machines, operating systems and compilers. This places additional burden on the science

data processing algorithms to be written as "generically" as possible, using a widely accepted language. An approach to ensuring future portability is to test the algorithm code on multiple platforms at acceptance time. If a program demonstrates portability across multiple computer platforms, it is more likely to be portable to new operating system versions or compilers in the future.

Even so, successful porting of the algorithms to a new system cannot be assumed. As a result, each migration must be accompanied by a significant processing campaign to verify that the algorithm produces the same results as before, using quality checks of the type described in an earlier subsection.

DERIVED DATA LATENCY

Unless sufficient computing resources are available, data delivery time could increase drastically over what is required to merely retrieve the needed data that has been derived *a priori*. However, note that the processing time latency problem may be counterbalanced by an increased ability to stage, or even keep online, the Level 0 data needed for the processing. Furthermore, the latency problem can be mitigated for real-time or quasi-operational use of the data using the short-term archive discussed briefly above, which would provide ready access to these data. Once these data are no longer available on the short-term archive, users might have to be granted varying levels of priority for assignment of computational resources.

CONCEPT OF OPERATIONS

To gain better understanding on how virtual products would work in an intelligent archive, a user scenario is presented below. The "user" could be, and for the case of the intelligent archive often is, an application (such as an assimilation model) rather than a person.

In this scenario, the intelligent archive is tasked by a user to provide derived parameters from the "best" algorithm. In this case, "best" really means most appropriate for the user's circumstances. This judgment would be made based on criteria such as:

- desired latency, which may prevent incorporation of some ancillary data
- desired science quality, which may not support execution until a later time or date due to availability of ancillary or sufficient data
- science-based vs. machine-learning based algorithms.

Although the definition of "best" will be purposely left vague, we can make the assumption that the intelligent archive system has some *a priori* rules that define how the best quality derived data will be produced on basis of available data, location and/or geography of the location of interest, seasonal considerations, and the accepted "best" algorithm for these conditions.

1. The user submits a request for a data product over a specific geographic region and time period. The size of the region need not be constrained to some minimum pre-defined "granule" size (unless required for processing quality checks). The user may specify latency or quality criteria for selecting the "best" available algorithm to generate the requested data.
2. The intelligent archive determines which of the "approved" algorithms provides the best retrieval on the basis of available data, location/geography, time/date/season, environmental conditions, etc. A variation of this might be an ensemble method where parameters are derived based on several algorithms and mean, confidence, and range of uncertainty are calculated and provided.
3. Before retrieving any data from the long-term archive, the intelligent archive would check a short-term storage area for data matching the user's needs. If available, the user will be sent notification that the data request is complete.
4. If the request data are not available in the short-term storage, the Level 0 data and associated metadata are retrieved, processed, and quality checked. The processed data are then placed in the short-term storage and the user is notified and provided information needed to retrieve the data.
5. The user then retrieves the processed data from the short-term archive.

CONCEPTUAL ARCHITECTURE

In order to look at the conceptual architecture, it is helpful to first examine how Virtual Data Products fit into the overall concept of an Intelligent Archive. Figure 1 shows the an adapted view of the Intelligent Archive concept envisioned by Ramapriyan et al[1]. Virtual data products are primarily a means of accomplishing the functions of Data Management and Data Persistence/Preservation Management. In doing so, they clearly need also to task the processing systems of Distributed Intelligent Systems. However, one area where virtual data products may be particularly useful is within Intelligent Sensors. Depending on the relative cost (and for spaceborne systems, radiation hardening) of computing elements vs. storage elements, virtual data products may be used to constitute a virtual archive on the sensor itself: the sensor might store just compressed, raw data in fairly

limited space, responding to requests for data products with on-the-fly productions of them.

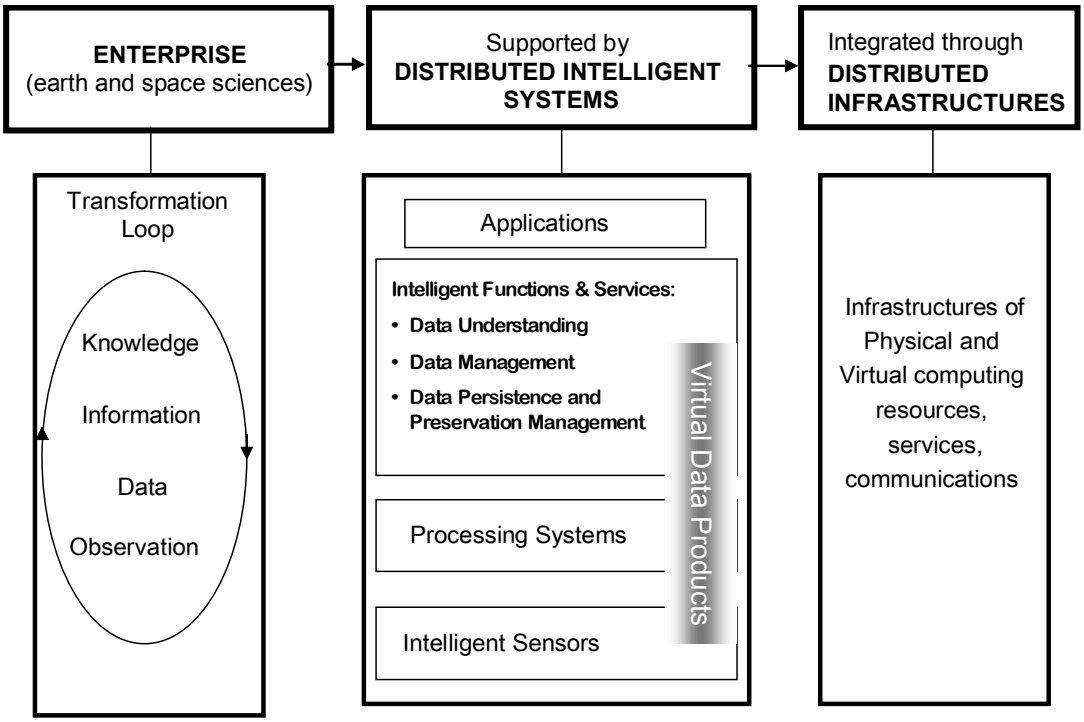

Figure 1. Virtual Data Products may be components of an Intelligent Archive's Data Management and Data Persistence/Preservation Management; they likely make use of the Processing Systems; and they may also provide a means to implement a virtual archive within Intelligent Sensors.

Figure 2 shows a possible architectural concept for the component responsible for Virtual Data Products. Three main aspects of the operation need to be considered to accommodate the scenarios discussed above.

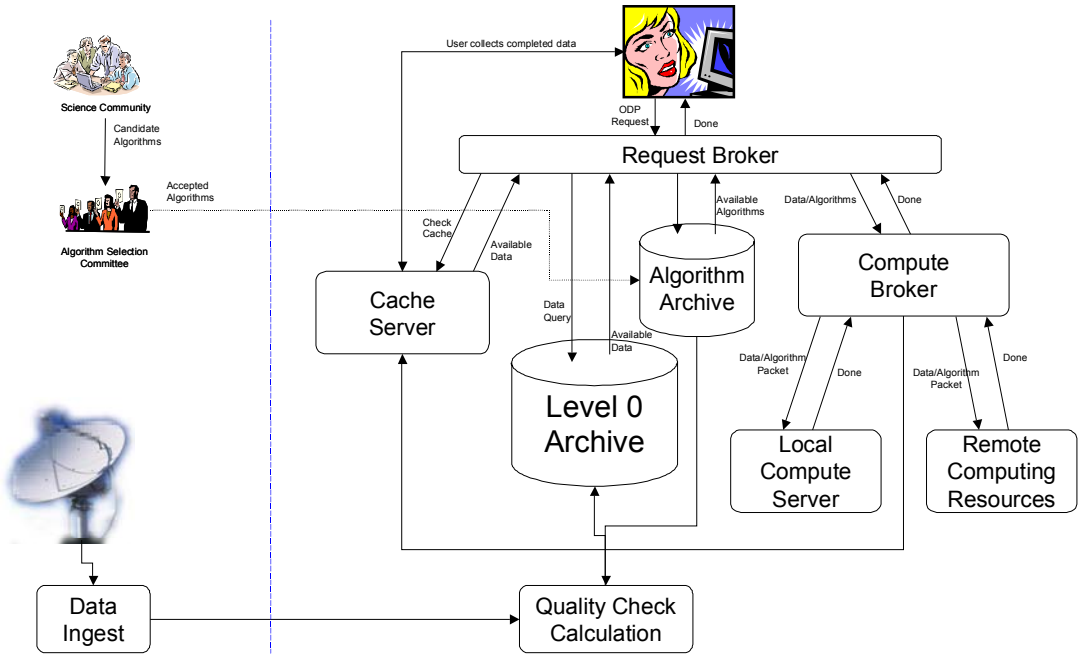

Figure 2 Conceptual Architecture for Processing Virtual Data

DATA PREPARATION AND STAGING

- Ingest: Although not part of the intelligent archive, the data receive and ingest functions are shown to illustrate the source of the data and its path into the intelligent archive
- Data Preparation: The most important aspect of this function is the Quality Check Calculations discussed earlier. This function is performed on incoming data using the available algorithms in the algorithm archive. Additionally, any cataloging and database preparation could be done here as well before committing these data to the Level 0 Archive.
- Staging: Initially, the intelligent archive should "stage" newly receive data to quick-access mass storage, or Cache Server, for some period to allow immediate access after initial acquisition without waiting for long-term storage retrieval. The amount of time to hold in mass storage could be based on typical lag time between collection and request.

DATA REQUEST/BROKERING/PROCESSING

REQUEST BROKERING

The potential complexities of available products, algorithms / versions, ancillary data, etc. are likely to require a component of the system that can present options and accept requests.

- The Request Broker allows users to interact with the intelligent archive. It is here, where users would see lists of available data, algorithms, and (depending on the quality check option used) thumbnail views of Level 1 and/or Level 2 data.
- When the user is satisfied with the request, the Request Broker accepts the users' request.

- The Request Broker then searches the Cache Server for available data. If it is not available, the Request Broker will pull the required data from the Level 0 Archive and appropriate algorithms from the Algorithm Archive. This bundled data/algorithm package is then passed to the Compute Broker for the computation of the required variables.

REQUEST PROCESSING AND ASSEMBLY

- The Compute Broker receives the bundled data/algorithm package from the Request Broker. Based on the user's priority, it may pass the bundle to a Local Compute Server (for immediate processing) or farm it out to Remote Computing Resources (for delayed processing as resources are available).
- The Local Compute Server might be a local high performance computing system or a local grid system. Remote Computing Resources could be a widely dispersed grid computing system. In either case, it is possible that the processing job could be shared between multiple computers by dividing the data/algorithm bundle into segments.
- Upon completion of all processing, the Compute Broker will re-assemble the processed data (if needed) and store it on the Cache Server. The Compute Broker will then notify the Request Broker of the completed request and the location of the processed data.
- Request Delivery: When notified by the Compute Broker of a successful completion of required computing, the Request Broker will then notify the user that the data retrieval is complete, and provide appropriate information for the user to retrieve the data from the Cache Server.

ALGORITHM DEVELOPMENT/ACCEPTANCE

Although this function is not strictly part of the intelligent archive, it certainly impacts the intelligent archive and its operation.

- The user community would submit algorithms with a sample data set and the resulting product, perhaps with the results of a standard test (designed by the committee) included.
- Members of the committee might include representatives from the user community, science team members, and the intelligent archive.
- Once an algorithm is accepted, it is entered into the Algorithm Archive. A processing campaign would then be initiated to execute the Quality Check Calculation for the new algorithms. Older algorithms (and associated quality check data) may be purged or retained depending on the confidence in the overall quality of the new algorithms.
- Data provided by one of these algorithms will also have an electronic signature or "seal of approval" to indicate that these are the best available data for studies. This signature will be needed to confirm and track the data "parentage."

CONCLUSION

The replacement of stored products, processed routinely, by virtual products processed on demand presents a number of challenges. Aside from the basic mechanics of implementing an on-demand processing system, issues of computing power, latency, quality assurance, and long-term migration to new technologies must all be addressed. However, all of these issues also have potential solutions that appear at least feasible within the next decade, given expected improvements in computing technology. In return, product virtualization would reduce storage, and more importantly data migration, requirements and provide user benefits in algorithm consistency and selection, as well as more precise data selection. If we neglect these latter benefits, perhaps the key factor on how much to invest in product virtualization is the relative growth in storage vs. computing technology. If computing technology leaves storage technology behind in the coming years, product virtualization will be easier to accomplish, and will likely become more necessary as data growth is likely to keep up with computing growth[1]. On the other hand, if storage technology comes out ahead, the driver for virtual products is lessened. Even if overall trends follow the latter case, it may still be important to examine specific environments, such as spaceborne data systems perhaps, where computing may be easier to deploy than storage.

---

[1] Ramapriyan, H. K., G. McConaughy, C. Lynnes, S. Kempler, K. McDonald, R. Harberts, L. Roelofs, and P. Baker, 2002. "Conceptual Study of Intelligent Archives of the Future", Report prepared for the Intelligent Data Understanding program, 39 p., http://daac/IDA/IA_report_8-27-02_baseline.pdf..

[2] Davis, F., W. Farrell, J. Gray, C. R. Mechoso, R. Moore, and M. Stonebraker, 1994. "EOSDIS Alternative Architecture- Final Report ", #ECS-00012, UC Berkeley ERL. Sept, 1994. Available online at http://research.microsoft.com/~gray/EOS_DIS/.

[3] Dao, S., and J. Feldman, 1995. "IAS Evaluation Document", ECS White Paper, Available online at: http://edhs1.gsfc.nasa.gov/waisdata/docsw/pdf/wp1700101.pdf.

[4] Deelman, E., C. Kesselman, G. Mehta, L. Meshkat, L. Pearlman, K. Blackburn, P. Ehrens, A. Lazzarini, R. Williams, and S. Koranda, "GryPhyN and LIGO, Building a Virtual Data Grid for Gravitational Wave Scientists", Available online at: http://www.isi.edu/~deelman/hpdc11.pdf.

[5] Botts, M., K. Keiser, H. Conover, S. Graves, 2000. "Visualization of On-demand Virtual Data Products in a Distributed Environment", submitted to International Cartographic Association Commission on Maps and the Internet Symposium 2000, Knoxville, Tennessee, USA, Oct. 11, 2000. Available online at: http://pm-esip.msfc.nasa.gov/library/uah_ica.pdf.

[6] Halem, M., F. Shaffer, N. Palm, E. Salmon, S. Raghavan* , and L. Kempster, 1999. "Technology Assessment of High Capacity Data Storage Systems: Can We Avoid A Data Survivability Crisis?" Greenbelt, MD: Earth and Space Data Computing Division, NASA Goddard Space Flight Center. Available online at: http://esdcd.gsfc.nasa.gov/ESDCD/whitepaper.data_survive.html.

[7] Esaias, W. E., Abbott, M. R., Barton, I., Brown, O. B., Campbell, J. W., Carder, K. L., Clark, D. K., Evans, R. H., Hoge, F. E., Gordon, H. R., Balch, W. M., Letelier, R., & Minnett, P. J., 1998. An overview of MODIS capabilities for ocean science observations. IEEE Transactions on Geoscience and Remote Sensing, 36(4), 1250-1265.